УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
Том 150, кн. 1 Физико-математические науки *2008*

УДК 004.93-021.64

РАЗЛИЧЕНИЕ ДВУХ ДИКТОРОВ ПО КОРОТКИМ ФРАЗАМ МЕТОДАМИ ОБРАБОТКИ ИЗОБРАЖЕНИЙ

Е.Л. Столов

Аннотация

В работе предлагается алгоритм, позволяющий различать двух дикторов по произнесенным фразам. Имеется набор коротких звуковых файлов, принадлежащих двум дикторам. Среди этих файлов имеются два F, G, о которых известно, что они принадлежат разным дикторам. Не предполагается, что указанные файлы соответствуют одной и той же фразе. Требуется определить принадлежность остальных файлов из множества. Задача решается с помощью аппроксимации фрагментов спектрограммы вейвлетами специального вида с последующей обработкой нейронной сетью. Приведены результаты экспериментов с файлами из звуковой базы.

Ключевые слова: различение дикторов, короткие фразы, графический метод.

Введение

Идентификация диктора по произнесенной речи остается весьма актуальной задачей. Основная проблема идентификации (аутентификации) диктора заключается в том, что непосредственное сравнение звукового файла с образцом не имеет смысла. Вследствие этого необходим подсчет по заданному файлу определенных статистик, а также метод принятия решения на основе вычисленных значений параметров. В свою очередь, подсчет статистик возможен после выбора какой-либо модели, согласно которой продуцируется речь. Обычно речевой сигнал рассматривается как реализация некоторого случайного процесса, и задача сводится к оценке параметров этого процесса. Принятие решения осуществляется на основе значений некоторой функции от найденных параметров. В качестве такой функции часто используется нейронная сеть. Как правило, отдельно рассматривают задачи идентификации диктора по продолжительной речи и по короткой фразе. В первом случае имеется большой набор данных для анализа, в связи с чем можно использовать сложные математические модели. Например, в задачах распознавания речи основной моделью стала модель сигнала на основе скрытых марковских цепей. Оценка параметров такой модели в настоящее время разработана достаточно хорошо, поэтому и для задачи идентификации диктора был использован тот же подход. Реализация этой идеи представлена в [1]. Однако основной моделью, применяемой для идентификации диктора, стала модель, согласно которой распределение амплитуды импульсов речевого сигнала может быть описано с помощью смеси гауссовских распределений [2]. Согласно [2] выборочная плотность $p(\bar{x})$ всего сигнала либо отдельных полос, выделенных гребенкой фильтров, аппроксимируется смесью

$$p(\bar{x}) = \sum_k b_k g_k(\bar{x}),$$

где $g_k(\bar{x})$ – плотность нормального распределения некоторой размерности D, $\sum\limits_k b_k = 1$. Идентификация диктора сводится к вычислению параметров смеси: подсчету коэффициентов и оценке среднего и ковариационной матрицы для каждого

из распределений. В основополагающей работе [2] в качестве значений параметров берутся оценки максимального правдоподобия. После этого появились многочисленные статьи, в которых предложены различные методы оценки параметров смеси и решающие правила для идентификации.

Иная ситуация складывается в случае, когда исходным материалом является небольшой фрагмент, состоящий из одной или нескольких коротких фраз. Очевидно, что достигаемая достоверность результата в этом случае значительно ниже. В этом ряду отдельно стоит задача идентификации диктора по парольной фразе, когда известен произносимый текст. Здесь предполагается наличие определенной статистики, поскольку фраза произносится многократно. Так, в работе [3] вводится понятие особой точки в звуковом файле, отвечающем парольной фразе, и изучается распределение особых точек в файле. Для идентификации диктора по коротким фразам предлагается также использовать анализ файла на уровне отдельных фонем. В работе [4], например, предлагается модель, где для оценки распределения амплитуд в гласных звуках применяется многомерное гауссовское распределение. В работе [5] используется нейронная сеть, которая тренируется для разделения дикторов на основе анализа произнесения отдельных фонем. В работе [6] предлагается более полный набор характеристик звукового файла, включающий спектральные и кепстральные коэффициенты. К этим значениям применяется метод главных компонентов, после чего решение принимается на основе анализа векторов малой размерности.

Несмотря на обилие подходов, задачу идентификации диктора по коротким фразам нельзя считать решенной. Это связано с принципиальными трудностями, возникающими при извлечении характерных параметров из-за недостаточности тестового материала. Например, анализ на основе фонем обладает существенным недостатком: отсутствует объективный критерий определения границ файла, отвечающего данной фонеме. Вследствие этого, такие границы при разных подходах к выделению оказываются разными, что затрудняет практическое использование указанного подхода. В этой связи представляют интерес методики, свободные от указанного недостатка.

В данной работе решается более простая задача. Имеется набор коротких звуковых файлов, принадлежащих двум различным дикторам. Среди этих файлов имеются два, про которые известно, что они принадлежат разным дикторам. Не предполагается, что указанные файлы соответствуют одной и той же фразе. Требуется определить принадлежность остальных файлов из множества. Идея предлагаемого алгоритма состоит в следующем. Звуковой файл разбивается на фрагменты. По каждому фрагменту строится спектрограмма, являющаяся поверхностью. Полученная поверхность аппроксимируется некоторым стандартным образом, в результате чего эта поверхность определяется малым числом параметров. После этого строится нейронная сеть, которая по вычисленным векторам различает поверхности, построенные по соответствующим файлам. Для этого сеть тренируется на известных файлах так, чтобы она принимала значение -1 на фрагментах файла первого диктора и 1 на фрагментах файла второго диктора. Наконец, все остальные файлы из набора классифицируются в зависимости от значений, найденных нейронной сетью для этих файлов. Реализация алгоритма в рамках представленной идеи предполагает выполнение дополнительных условий: результаты не должны зависеть от коэффициента усиления аппаратуры; влияние фонового шума должно быть сведено к минимуму; существует возможность учета влияния полосы пропускания канала связи, использованного для записи файла. Очевидно, что на практике трудно выполнить все указанные условия, а излагаемая ниже методика лишь в какой-то мере удовлетворяет отмеченным требованиям.

1. Построение спектрограммы и ее сжатое описание

Как отмечалось выше, конечное решение о принадлежности файла одному из дикторов принимается на основе значений некоторой нейронной сети. Нейронная сеть имеет стандартную структуру [8]. Если на вход сети поступает некоторый вектор-столбец v, то на вход нейрона с номером k подается сигнал вида $s_k = w_k * v + a_k$. Здесь вектор-строка w_k и число a_k меняются в процессе тренировки сети. Нейрон с номером k преобразует полученный сигнал в число $u_k = f(s_k)$, где f – некоторая функция. Линейная свертка значений u_k, выработанных нейронами одного слоя, подается на выход сети либо на вход нейрона следующего слоя. При этом коэффициенты свертки также меняются в процессе тренировки. Каждый звуковой файл разбивается на отдельные перекрывающиеся фрагменты, находятся спектры этих фрагментов, а на вход сети подаются векторы, характеризующие спектральные свойства каждого из найденных фрагментов. Поскольку процедура распознавания ориентирована на файлы малой длины, число указанных векторов также будет небольшим. Для того чтобы в этом случае тренировка нейронной сети была корректной, следует ограничить размерность этих векторов так, чтобы число векторов значительно превышало их размерность. В противном случае использование нейронной сети теряет смысл, поскольку, как следует из определения используемой сети, дело сведется к построению линейной дискриминантной функции.

Перейдем к алгоритму построения векторов, аппроксимирующему спектр фрагмента. Все фрагменты имеют одну и ту же длину L. Выбирается окно длины W, применяемое для подсчета преобразования Фурье от сигнала. Окно движется вдоль фрагмента, сдвигаясь каждый раз на половину своей длины. Для каждого положения окна подсчитываются коэффициенты Фурье от вектора, попавшего внутрь окна. Поскольку предполагается использовать быструю схему для вычисления преобразования, число W выбирается равным степени 2. В результате описанной процедуры фрагмент заменяется последовательностью векторов – наборов коэффициентов Фурье, отвечающих окнам внутри фрагмента. Это и есть спектрограмма фрагмента. Поскольку исходный сигнал вещественный и в дальнейшем используются только модули коэффициентов Фурье, принимаются во внимание лишь первые $W/2+1$ из найденных значений. Описание полученной спектрограммы фрагмента в виде набора найденных коэффициентов Фурье требует задания большого числа значений, поэтому оно не пригодно для непосредственной подачи на вход нейронной сети. Понижение размерности этого описания осуществляется в несколько этапов. На первом этапе применяется таблица BARK [7]. Суть этой таблицы заключается в следующем. Согласно физиологическим наблюдениям, человеческое ухо не различает частоты, находящиеся внутри определенных полос звукового спектра. Таблица BARK задает границы этих полос. Очевидно, что эта таблица определяет некоторые усредненные значения. Понижение размерности достигается заменой всех коэффициентов, отвечающих частотам из одной полосы, их суммой квадратов модулей (энергией). Поскольку человек, как правило, может различить дикторов, использование указанной таблицы для понижения размерности не должно существенно исказить конечный результат. Это значительно понижает размерность рассматриваемых векторов. Кроме того, предложенная процедура позволяет при желании оставить лишь нужную полосу спектра для дальнейшего исследования. Необходимость такого урезания спектра возникает, если исходный сигнал получен по каналу с ограниченной шириной пропускания. Границы полос приведены ниже в табл. 1. Здесь верхняя строка содержит номера полос, а строка под ней – частоту, с которой начинается полоса и заканчивается предыдущая полоса.

Табл. 1

Полосы частот таблицы BARK

Номер полосы	1	2	3	4	5	6	7	8
Начало полосы	100	200	300	400	510	630	770	920
Номер полосы	9	10	11	12	13	14	15	16
Начало полосы	1080	1270	1480	1720	2000	2320	2700	3150
Номер полосы	17	18	19	20	21	22	23	24
Начало полосы	3700	4400	5300	6400	7700	9500	12000	15500

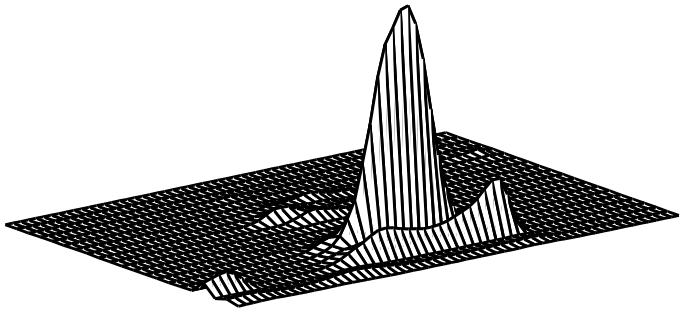

Рис. 1. Фрагмент спектра, сжатого с учетом таблицы BARK

Используя номер полосы из таблицы в качестве параметра, получаем поверхность, где по оси X откладывается положение окна, с помощью которого подсчитывается преобразование Фурье, по оси Y откладывается номер полосы из таблицы BARK, а по оси Z – сумма квадратов модулей коэффициентов Фурье, попавших в одну полосу таблицы. Пример такой поверхности представлен на рис. 1. В дальнейшем будем говорить, что поверхность представлена матрицей, элемент $m_{i,j}$ которой определяет точку в пространстве с координатами $i,\ j,\ m_{i,j}$.

Следующий этап сжатия является наиболее трудоемким, и для его реализации применимы параллельные вычисления. Этот этап заключается в аппроксимации поверхности, найденной на предыдущем шаге, с помощью неортогональных вейвлетов специального вида. Процедура аппроксимация подробно описана в [9], поэтому здесь дадим лишь минимальные сведения, необходимые для понимания дальнейшего текста. Каждый вейвлет определяется матрицей ранга 1 и положением центра. Поверхность, отвечающая матрице вейвлета, имеет вид, представленный на рис. 2. Спектрограмма фрагмента аппроксимируется суммой вида

$$P = \sum_{s=1}^{R} c_s P_s, \tag{1}$$

Здесь каждое слагаемое задается размерами матрицы, положением центра поверхности и коэффициентом c_s, с которым слагаемое входит в сумму. Другими словами, каждое слагаемое в (1) определяется пятью числами. Пример поверхности, полученной в результате аппроксимации поверхности на рис. 1, представлен на рис. 3. Исходная поверхность имеет носитель, состоящий из 15×60 точек, а для аппроксимации поверхности использованы 10 слагаемых. Способ распараллеливания алгоритма для отыскания разложения (1) представлен в [9]. Таким образом, получается сжатое описание спектрограммы фрагмента. Нейронная сеть тренируется так, чтобы она различала фрагменты, отвечающие разным дикторам.

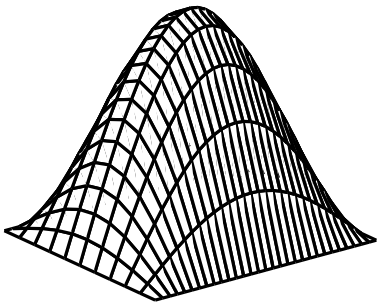

Рис. 2. Вейвлет, применяемый для аппроксимации поверхности

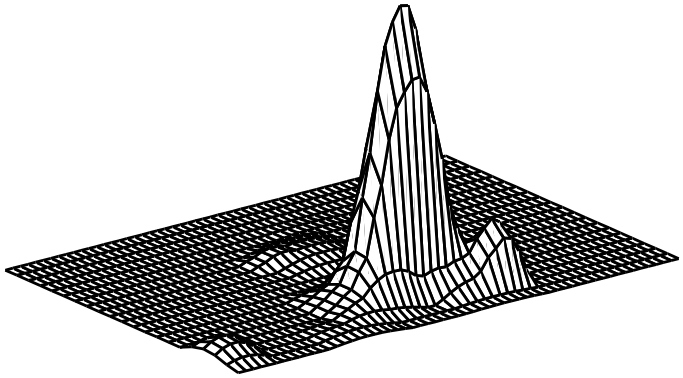

Рис. 3. Результат аппроксимации поверхности суммой вейвлетов

Следует дать обоснование применяемому методу сжатия спектрограмм. Основная проблема, связанная с использованием спектрограмм для идентификации дикторов, заключается в случайном выборе начала фрагментов. Если процедуру образования фрагментов применим к одному и тому же звуковому файлу, поменяв лишь начало отсчета, в результате получим разные наборы спектрограмм. В этой связи основная цель, преследуемая в процедуре сжатия, заключается в том, чтобы сделать сжатые образы менее чувствительными к положению начала отсчета. В случае стационарного сигнала сдвиг окна не влияет на модули коэффициентов Фурье. Это компенсирует малые отклонения в начале отсчета. Аппроксимации, согласно (1), двух одинаковых поверхностей, имеющих разные начала отсчета, различаются сдвигом центров слагаемых по координате X на одно и то же число. Этот сдвиг легко компенсируется в процессе подготовки данных для нейронной сети, о чем будет сказано ниже.

2. Результаты экспериментов

Исходным материалом для экспериментов послужили звуковые файлы из базы TIMIT. Файлы записаны с частотой 16 kHz по два байта на отсчет. Для каждого диктора имеется набор из 10 файлов. Ниже приведены результаты одного из проведенных экспериментов, которые типичны для рассматриваемых наборов. Среди фраз, записанных женскими голосами, имелись две одинаковые фразы, а остальные фразы в наборах не совпадали. Для сжатия спектрограмм использовались окна длиной в 128 отсчетов. Из полученных коэффициентов сохранялись

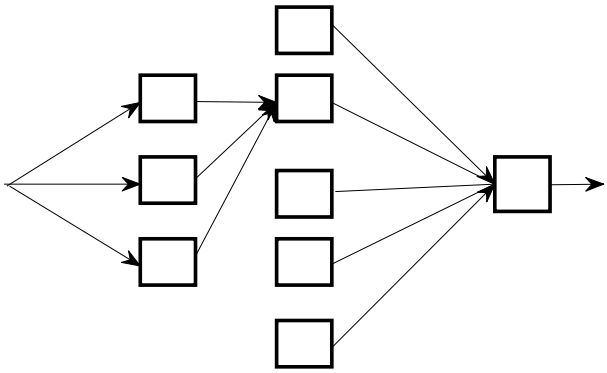

Рис. 4. Нейронная сеть

Табл. 2

Результаты эксперимента

Файл	F_1	F_2	F_3^+	F_4^+	F_5^*	F_6^+	F_7	F_8^+	F_9^+	F_{10}^+
Длина (kb)	93	69	98	90	114	50	56	47	72	78
Отрицательные	94	63	105	102	234	50	43	48	79	86
Положительные	95	76	101	86	15	41	59	36	64	72
Файл	G_1^+	G_2^+	G_3^*	G_4	G_5^+	G_6	G_7^+	G_8^+	G_9	G_{10}
Длина (kb)	104	81	115	81	51	91	64	81	72	88
Отрицательные	108	82	17	95	47	100	60	86	109	93
Положительные	109	90	233	66	48	95	69	91	77	74

модули лишь первых 65 коэффициентов, которые после сжатия, согласно таблице BARK, превращались в 15 чисел (использовались полосы с номерами от 3 до 17). Спектрограмма фрагмента строилась по 20 положениям окна. Затем поверхность аппроксимировалась согласно (1) для $R = 10$, а число строк и столбцов в матрице каждого слагаемого не превышало 11. Другими словами, фрагмент кодировался вектором длины 50. Однако в таком виде эти векторы не пригодны для тренировки сети. Прежде всего следовало учесть смещение начала отсчета. С этой целью находилось слагаемое в (1), для которого координата X имела наименьшее значение. Этот минимум вычитался из всех X-координат остальных слагаемых. Таким образом компенсировался сдвиг начала отсчета, а длина вектора фрагмента уменьшалась на единицу. Координаты этого вектора имеют разную природу (размерность матрицы, положение центра, коэффициент), вследствие чего диапазоны изменения каждой из координат сильно различаются. В то же время для успешного проведения процесса тренировки нейронной сети желательно, чтобы эти диапазоны были близки. Для достижения указанного эффекта компоненты векторов нормировались следующим стандартным образом. Пусть $z_{m,n}$ – компонента с номером n в векторе, отвечающем фрагменту с номером m. Подсчитывалось число $M_n = \max_m |z_{m,n}|$. После этого число $z_{m,n}$ заменялось на $z_{m,n}/M_n$. В результате диапазон изменения каждой компоненты превращался в отрезок $[-1, 1]$. Нейронная сеть состоит из 3 слоев, содержащих 3, 5 и 1 нейронов соответственно. Нейроны в первых двух слоях реализуют гиперболический тангенс, а на последнем слое стоит обычный сумматор. Условная схема сети представлена на рис. 4. Условность рисунка заключается в том, что на самом деле сигнал с каждого нейрона первого слоя поступает на вход каждого нейрона второго слоя. Выбор указанной

конфигурации ни в коем случае не является обязательным. В данной работе конфигурация была найдена в результате многочисленных экспериментов. Из каждого набора для тренировки сети было выбрано по одному файлу, имеющему примерно одинаковую длину. Файлы соответствуют разным фразам. Сеть тренировалась так, чтобы на фрагментах первого диктора значения на выходе равнялись -1, а на фрагментах второго диктора были равны 1. Решающее правило для идентификации диктора выглядит следующим образом. Все спектрограммы звукового файла в сжатой форме подаются на вход сети, и подсчитывается число отрицательных и положительных значений на выходе. Если число отрицательных значений превышает число положительных значений, файл помечается как принадлежащий первому диктору, в противном случае считается, что он принадлежит второму диктору. Все результаты сведены в табл. 2. В таблице знаком "+" отмечены правильно идентифицированные файлы, а знаком "*" отмечены файлы, использованные для тренировки сети. Невысокая достоверность обнаружения объясняется малой длиной файлов, использованных в экспериментах. Нетрудно видеть, что файлы, задействованные для тренировки сети, имеют продолжительность чуть менее 4 с. Существующие системы идентификации требуют, чтобы продолжительность речи для настройки измерялась несколькими минутами.

3. Выводы

Проведенные эксперименты показывают принципиальную возможность использования предложенной методики для идентификации дикторов. В тех случаях, когда отсутствует тестовый материал подходящей длины, можно использовать предлагаемую методику. Достоверность идентификации, обеспечиваемой рассмотренным методом, требует дальнейшего исследования.

Summary

E.L. Stolov. Speaker Identification by Short Phrases Using Image Processing Procedure.

A new algorithm for speaker identification is suggested. A set of sound files belonging to two speakers is given. There are two short files known to be corresponding to two different speakers. The speakers are not supposed to have articulated the same phrase. The task is to establish belonging of each file in the set. The problem is solved using an approximation of spectral surface of sound file by wavelets of special kind. The compressed form of the spectral surface is processed by a neuron net. The decision about the belonging is made basing on the values produced by the neuron net. Some results of an experiment with files from a speech database are presented.

Key words: speaker distinguishing, short phrases, graphical method.

Литература

1. *Rosenberg A.E., Lee C.-H., Soong F.K.* Sub-word unit talker verification using hidden markov models // Proc. ICASSP. – 1990. – P. 269–272.
2. *Reynolds D.A., Rose R.C.* Robust text-independent speaker identification using gaussian mixture speaker models // IEEE Trans. Speech and Audio Processing. – 1995. – V. 3.– P. 72–83.
3. *Столов Е.Л.* Идентификация диктора на основе отыскания особых точек в произнесенной фразе // Вестн. Томск. гос. ун-та. Приложение. – 2006. – № 17. – С. 37–40.
4. *Zilca R.D.* Text-Independent Speaker verification using utterance level scoring and covariance modeling // IEEE Trans. Speech and Audio Processing. – 2002. – V. 10. – P. 363–370.

5. *Liou H.-Sh., Mammone R.J.* Speaker verification using phoneme-based neural tree networks and phonetic weighting scoring method // Proc. of the 1995 IEEE Workshop. – 1995. – P. 213–222.

6. *Magrin-Chagnolleau I., Durou G., Bimbo F.* Application of time-frequency principal component analysis to text-independent speaker identification // IEEE Trans. Speech and Audio Processing. – 2002. – V. 10. – P. 371–378.

7. *Huang X., Acero A., Hon H.-W.* Spoken language processing: A Guide to theory, algorithm, and system development. – New Jersey: Prentice-Hall, 2001. – 965 p.

8. *Gupta M.M., Liang Jin, Homma N.* Static and Dynamic Neural Networks: From Fundamentals to Advanced Theory. – Hoboken, NJ: Wiley-IEEE Press, 2003. – 752 p.

9. *Столов Е.Л., Шляпников А.В.*Распознавание лиц на фотографии путем анализа характерных областей // Учен. зап. Казан. ун-та. Сер. Физ.-матем. науки. – 2007. – Т. 149, кн. 2. – С. 138–145.

Поступила в редакцию
26.12.07

Столов Евгений Львович – доктор технических наук, профессор кафедры системного анализа и информационных технологий Казанского государственного университета.
E-mail: *Yevgeni.Stolov@ksu.ru*